In Quest for Requirements Engineering Oracles: Dependent Variables and Measurements for (good) RE

In Quest for Requirements Engineering Oracles: Dependent Variables and Measurements for (good) RE

Daniel Méndez Fernández, Jakob Mund, Henning Femmer, Antonio Vetrò
Technische Universität München, Germany
http://www4.in.tum.de/~mendezfe|femmer|mund|vetro

ABSTRACT

Context: For many years, researchers and practitioners have been proposing various methods and approaches to Requirements Engineering (RE). Those contributions remain, however, too often on the level of apodictic discussions without having proper knowledge about the practical problems they propagate to address, or how to measure the success of the contributions when applying them in practical contexts. While the scientific impact of research might not be threatened, the practical impact of the contributions is. **Aim:** We aim at better understanding practically relevant variables in RE, how those variables relate to each other, and to what extent we can measure those variables. This allows for the establishment of generalisable improvement goals, and the measurement of success of solution proposals. **Method:** We establish a first empirical basis of dependent variables in RE and means for their measurement. We classify the variables according to their dimension (e.g. RE, company, SW project), their measurability, and their actionability. **Results:** We reveal 93 variables with 167 dependencies of which a large subset is measurable directly in RE while further variables remain unmeasurable or have too complex dependencies for reliable measurements. We critically reflect on the results and show direct implications for research in the field of RE. **Conclusion:** We discuss a variety of conclusions we can draw from our results. For example, we show a set of first improvement goals directly usable for evidence-based RE research such as "increase flexibility in the RE process", we discuss suitable study types, and, finally, we can underpin the importance of replication studies to obtain generalisability.

Categories and Subject Descriptors

D.2.1 [**Software Engineering**]: Requirements/Specification

General Terms

Measurement, Experimentation

EASE '14, May 13-14 2014, London, UK

Keywords

Evidence-based Research, Requirements Engineering, Metrics and Measurements

1. INTRODUCTION

Requirements Engineering (RE) aims at the discovery and specification of requirements that unambiguously reflect the purpose of a software system as well as the needs of all relevant stakeholders. The specification of precise requirements directly contribute to appropriateness and cost-effectiveness in the development of a system [21] and, thus, RE is an important factor for productivity and (product) quality [5]. Given the practical importance of RE, it remains an inherently complex discipline due to the various influences in practical environments making the process itself uncertain [17]. The interdisciplinary nature of the field and the dependency on the various human factors that pervade RE like no other software engineering discipline, eventually make RE hard to investigate and even harder to improve [16].

In response to the practically motivated relevance and the fundamental challenges given in the discipline, we can observe over the last two decades a strong research community arising from an initially neglected field of investigation. In the course of various research endeavours, a plethora of methods and approaches to RE have been proposed.

However, when it comes to evaluating the contributions in practical environments, which is a prerequisite for providing insights into factors relating to the practical impact of the contributions, we can observe that available evaluations are often not in tune with the (practical) problems they are intended to address. Available contributions provide, if at all, isolated case studies investigating aspects that hardly can be generalised, e.g., long-term views on cost and benefits when applying developed methods. And in most cases accurate evaluations starve in the future work section of the publication [2].

From an empirical perspective, however, the investigations one would expect with the contributions cannot be provided in short notice as:

- The effort necessary to conduct case and field study research is, in general, very high, and often can be portrait as an own scientific contribution itself.
- The accuracy and objectivity not only depend on the chosen study population, but also on the involved researchers. This threat implies that, ideally, the evaluations should be performed independently by researchers who are not involved in the development of the method under analysis. Given the current stress fields of aca-

DOI: http://dx.doi.org/10.1145/2601248.2601258

demic research environments, the independent investigation of given (methodological) contributions via confirmatory studies seems often to be unattractive.
- The external validity seems to demand for replication studies and longitudinal studies that are often not in scope of research projects.

Over and above all, investigations on selected benefits are often driven by very specific-context problems where we evaluate a solution using metrics and measurements that are important to a particular context whereas they might not be important to the other; for instance, supporting the creation of consistent RE artefacts might be important to one company while the focus of another company might be set on supporting communication within local teams.

It is thus not surprising that empirical evidence on the suitability of scientific contributions to tackle practical problems is per se difficult to provide and often underrepresented in scientific contributions. Another reason for the currently weak state of evidence in RE [2] might be the missing awareness of the RE community towards the possibilities we have in evidence-based research. Condori-Fernandez et al. [4] conducted a study to understand the extent to which the empirical software engineering research methods are adopted in the RE community. One result was that even senior researchers are not aware of the full potential of evidence-based research, not only to evaluate contributions on basis of hypotheses, but also to rigorously explore the problem domain itself to reveal theories and practically relevant improvement goals.

To develop solutions that are meant to solve commonly accepted problems, we first need to understand what commonly accepted problems are, how they evolve within the whole project ecosystem (beyond RE), and how we can infer a generalised notion of a "good RE" from those problems. This notion might not necessarily apply to all socio-economic contexts with individual and isolated problems, improvement goals, and organisational cultures, but it should support the reliable evaluation of solutions against a set of commonly accepted (dependent) variables and measurements. Having a set of variables and understanding to what extent those variables eventually matter from a practical perspective supports an evaluation of solution proposals and, as an ultimate goal, the establishment of RE oracles. For example, knowing which variables matter in practice and how they can be measured allows us to evaluate engineering methodologies w.r.t., e.g., their support to tackle communication flaws within teams or the creation of consistent artefacts as we know to what extent those problems matter and how they manifest themselves in the whole development process.

1.1 Problem Statement

To develop proper improvement goals, we first need an empirical basis of measurable variables that originate in RE and their dependency to further variables within the whole project ecosystem. A characterisation of such phenomena gives a better understanding of practical improvement goals, which allow, in turn, to develop and evaluate scientific contributions against practically relevant problems we are able to characterise and measure.

1.2 Research Objective

The main objective of the paper is to provide a first empirical basis of structured variables originating in RE and their dependency to further variables within the whole project ecosystems. A further classification of those variables according to their measurability and actionability allows finally to critically reflect on the possibilities in evidence-based research and to draw direct research implications for RE.

1.3 Contribution

In this paper, we transfer the results of a survey we conducted to identify a set of RE-related problems and their effects to a set of dependent variables in RE, show where those variables occur in the project ecosystem, and critically discuss the measurability of those variables. Based on those results and their critical discussion, we infer first research implications for RE research.

1.4 Outline

In Sect. 2, we discuss fundamentals and work related to our research. In particular, we introduce the *NaPiRE* project from which we infer the variables classified and analysed in context of this paper and introduce further research results on which we rely during our classification. In Sect. 3, we then introduce our overall research design. We present our results in Sect. 4.1. In Sect. 5, we critically reflect on our results and draw first research implications for evidence-based RE research in Sect. 6, before concluding with a discussion of the threats to validity and future work in Sect. 7.

2. FUNDAMENTALS AND RELATED WORK

Much work has been carried out to explore promising research directions in Requirements Engineering mostly baptised with the term "Roadmap". Prominent examples are the one by Nuseibeh et al. [21] and subsequent contribution by Chen et al. [2]. Both contributions explore the facets of RE in great detail and show various research directions within those facets; for example, regarding requirements elicitation techniques, modelling and analysis techniques, or aspects relating to RE as an interdisciplinary area which characterises the discipline in greatest distinction to other software engineering disciplines. However, when it comes to directly characterising RE more explicitly by its phenomena, which is necessary to, inter alia, infer measurable improvement goals for general RE research and to evaluate solution proposals against a common understanding of practically relevant RE variables, only few contributions are at our disposal.

Gorschek et al. [9] proposed a framework to characterise dependent variables in RE in context of a project ecosystem via five dimensions (from a process improvement perspective), namely:

1. *Requirements phase*
2. *Project* including variables like cost and time or project estimates
3. *Product* where dependent variables determine the degree of product success
4. *Company* considering the effects in multi-project environments or a product placed in a market
5. *Society*

In their contribution, they describe an initial set of dependent variables to support a view on the notion of RE quality in a broader context as within selected particularities of RE alone. However, the variables proposed so far were an initial set that needed further investigation and extension.

To reveal RE phenomena and characterise the discipline from a practical perspective, we mostly rely on (exploratory)

DOI: http://dx.doi.org/10.1145/2601248.2601258

field study and survey research. One of the most well known surveys is the Chaos Report of the Standish Group, examining, for example, root causes for project failures of which most are to be seen in RE, such as missing user involvement. Apart from having serious flaws in its design negatively affecting the validity of the results [7] studies of this type do not support investigation of contemporary phenomena and problems in industrial RE environments (as they focus on project failure only). Such investigations have, for example, been indirectly conducted by Damian et al. [5]. They analysed process improvements in RE and the relation to payoffs regarding, for example, productivity and the final product quality. Nikula et al. [20] present a survey on RE at organisational level of small and medium size companies in Finland. Based on their findings, they inferred improvement goals, e.g. on optimising knowledge transfer. A study used to infer recommendations to practitioners, such as user involvement in the elicitation process, has been performed by Enam et al. [6]. A more curiosity-driven study to analyse typical project situations in companies was presented by us in [17]. We could discover 31 project characteristics that directly influence RE. A survey that directly focused on discovering problems in practical settings was performed by Hall et al. [10]. They empirically underpin the problems discussed by Hsia et al. [11] and investigated a set of critical organisational and project-specific problems, such as communication problems, inappropriate skills or vague requirements, while those problems matched to a large extent project characteristics we could discover.

Still, studies as the mentioned ones, although being all valuable as they provided necessary groundwork for an empirical understanding of RE, had either their focus on RE phenomena without any relation to further phenomena in a project ecosystem (except project failure), or they were performed in isolation in one single company, thus, remaining not representative.

For this reason, we initiated a global family of surveys to reveal the status quo in industrial requirements engineering, namely the *NaPiRE* project ("Naming the Pain in Requirements Engineering") [16, 18]. This project provides, in the long run, an empirical survey repository and is performed in collaboration between various members of the *International Software Engineering Research Network* (ISERN). Under `http://re-survey.org`, we provide further information on the project as well as the complete survey data as it is published to the PROMISE repository.

For the purpose of the paper at hand, we rely on a subset of the NaPiRE data obtained from its run in Germany in 2013 with 58 companies [16, 18]. We take a set of problems in RE and their effects as revealed via open questions in NaPiRE and classify those variables using the scheme introduced by Gorschek et al. [9] (see also the next section).

3. RESEARCH DESIGN

Our aim is to provide a first empirical basis of structured variables originating in RE and their dependency to further variables within the whole project ecosystems. This allows for a critical reflection on metrics and measurements used for evidence-based RE research and the inference of research implications for RE research. To this end, we define three research questions as summarised in Tab. 1.

In RQ 1, we explore the variables and their manifestation in context of RE as they result from NaPiRE and transfer those variables to a subset of the dimensions defined by Gorschek at al. [9] (see also Sect. 2). We change the dimensions in response to disagreements during the classification procedure (e.g., we remove the dimension "product" as we see a product as a subset of artefacts occurring in multiple dimensions). The goal is to understand the problems named by practitioners in terms of where they occur and how they relate to each other.

Table 1: Research questions

RQ 1	Which RE-related phenomena exist, where in the project ecosystem do they manifest themselves, and how do they relate to each other?
RQ 2	Are the phenomena measurable?
RQ 3	Are the phenomena actionable?

To this end, two of the authors individually classified the RE phenomena according to the dimensions (*RE, Engineering, SW project, Company*), based on the individual, original statements that were given in the NaPiRE report. After this, each disagreement was discussed in depth until the researchers agreed on one common answer.

In RQ 2, we aim to identify those phenomena amenable to *measurement*. Therefore, a phenomenon is considered measurable [**?**] if and only if

(i) its understanding is sufficiently mature such that
(ii) an existing or anticipated measure, i.e., objective mapping to mathematical objects
(iii) can efficiently (e.g., in justifiable time) and
(iv) effectively (i.e., preserving empirical observations) capture the phenomenon
(v) under practical conditions and when applied on study objects which can be expected to be present in a software project ecosystem.

We further distinguish between artefacts or activities as the primary kind of study object, or both in cases where the phenomena can be measured on artefacts and activities as well. Please note that our intention is not to provide a taxonomy of measurements, but rather rate the extent of general measurability.

For the classification, we rely again on researcher triangulation. To this end, three authors of this paper (one senior researcher and two PhD students) classified the phenomena on a nominal scale (*not measurable, measurable on artefacts, measurable on activities, measurable on both*). The individual classifications lead to *full agreement*, *partial agreement* or *disagreement*. Also in this case, the three researchers discussed each disagreeing classification was discussed in depth until the researchers agreed on one common answer or at least a majority vote was possible.

In RQ 3, finally, we aim at investigating a certain property of the phenomena called *actionability* (see also [19]). An actionable phenomenon allows to make an empirically-informed decision based on its measurement with significant impact on the phenomenon, e.g., to improve/change the status of a phenomenon. Similar notions are interpretative guidelines or recommendations for action. For the identification, we rely again on the three authors who estimated the actionability of each phenomenon either by *yes* or *no*, leading to only *full* or *partial agreement* (majority vote) on this question.

DOI: http://dx.doi.org/10.1145/2601248.2601258

Validity Procedure

The three researchers who did the classification have several years of experience on RE. To analyse and increase the validity of the procedure, the classification was conducted independently by the three authors while we used the *kappa*-values to assess the agreement. Furthermore, we forced that the researchers discuss each disagreement in order to ensure a common understanding and reach a final classification where all the raters fully agreed or at least a majority vote was possible.

4. DEPENDENT VARIABLES

As a results of the survey, we revealed 93 variables with 167 dependencies. Given the complexity of the network, the complete list of all phenomena with the determined dimension of occurrence is available online[1].

Herein, we present only the most relevant relations between phenomena (also referred as variables). Fig. 1 illustrates those interconnected variables.

Each variable is represented as a box in the graph, labeled in the upper face with an unique alphanumerical identifier. The first part of the identifier is either "RP", i.e. requirement problem, or "M", i.e. manifestation of the problem, or "R", i.e. reasoning, and it is followed by a sequential number. For further explanations on problems, reasonings and manifestations, we point the reader to the available publications on the NaPiRE survey [16, 18]. The two front faces report the results of the classification. On the leftmost, the measurability (RQ2), on the rightmost the actionability (RQ3). As far the classification on dimensions is concerned, the nodes are placed in four different boxes, each one corresponding to a different dimension. The names of the variables presented in the graph are reported in Tab. 2.

The graph is directed: being A and B two connected phenomena, the relation A → B states that A causes B or manifests itself in B, according to participants' answers. The width of the arrow indicates how many respondents connected A to B: it can be interpreted, approximately, as the amount of evidence provided by the survey results on the relationship between A and B. The size of the node corresponds to the sum of the the weights of the outgoing and incoming connections. In the following, we present our results on finding dependent variables, structured according to the research questions.

4.1 RE-related Phenomena (RQ 1)

The classification of the variables according to dimensions resulted in 38 variables assigned to dimension *SW project*, 33 to *RE*, 14 to *Engineering*, and 8 to *Company*. The variables are highly connected (167 overall connections), but 50 % of nodes have only two connections. Especially the variables in the dimensions *SW project* and *RE* have frequent connections. We represent a portion of the graph in Fig. 1, selecting only those connections which are supported by more than one respondent (i.e., weight > 1, corresponding to 80% of all connections).

The most frequently observed relationship is *Moving targets* (RP02) → *Change requests* (M09). While this relationship might be obvious, it is interesting to notice that *Moving targets* manifests itself (outgoing edge), in this filtered graph, only with phenomena in the dimension *SW project*. On the other hand, the reasons for change requests (incoming edges) result from multiple dimensions. *Underspecified requirements* (RP05) and *Gold plating* (RP10), instead, affect only phenomena related to *SW project*.

The node with the highest frequency of incoming/outcoming connections is *Incomplete or hidden requirements* (RP01) in *RE*. Although it is affected only by one phenomenon, its manifestations are transversal to all dimensions and refer in particular to time delays and wasted efforts.

A separated network of dependencies, on the right, is built around the phenomenon *Additional Communication and Replaning* (M13), which, interestingly, is assumed to be directly caused only by phenomena from the RE dimension: *Terminological problems* (RP11), *Communications flaws to customer* (RP08), and *Inconsistent requirements* (RP07).

We observe yet another isolated network on the bottom of the RE dimension, around the node *Time boxing* (RP03), which is affected by three variables from *RE*, and *Bugs and defects* (M24) from *Engineering*. It is interesting to note that this is the only connection in which *Bugs and defects* appears (and in fact the size of the node is not large), which reveals that problems and bad quality in requirements might not propagate to the point to affect the external quality of code artefacts.

A final observation is that no circular dependencies (.i.e., chains of cause-effect relationships) are presented in this filtered graph.

4.2 Measurability of Variables (RQ 2)

Regarding RQ 2, we investigated which phenomenon, and thus potential variables, can be measured using an existing or anticipated metric (cf. Sec. 3), and on what study ob-

Table 2: Descriptions for variables in Fig. 1

ID	Name
M02	Underspecified Reqs.
M03	Incomplete Reqs.
M05	Effort and time overrun
M08	Failed approval of reqs.
M09	Change Requests
M13	Additional communication and replanning
M15	Failed Acceptance
M17	Increased effort in testing
M18	Increased effort in reviews
M20	Too complex solutions
M24	Bugs and defects
M29	Time overrun
M30	Cost overrun
M32	Stagnating progress
M36	Customer dissatisfaction
R01	Implicit Reqs not made explicit
R03	Missing abstraction from solution level
R11	Weak communication
R18	Too ambitious time planning
RP01	Incomplete / hidden reqs.
RP02	Moving targets
RP03	Time boxing
RP04	Separation reqs. from known solutions
RP05	Underspecified reqs.
RP06	Communication flaws in team
RP07	Inconsistent reqs.
RP08	Communication flaws to customer
RP10	Gold plating
RP11	Terminological problems

[1] `http://www4.in.tum.de/~mendezfe/openspace/NaPiRE/ease2014.zip`

DOI: http://dx.doi.org/10.1145/2601248.2601258

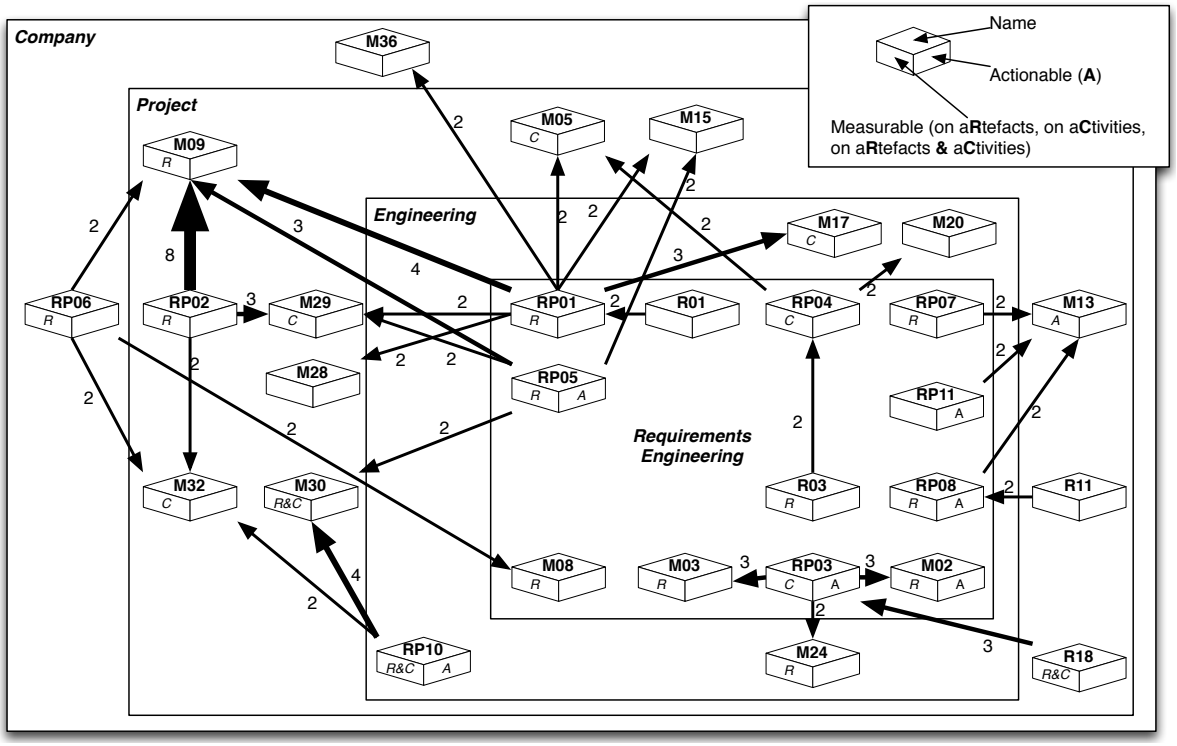

Figure 1: Relationship between RE-related variables with weight > 1.

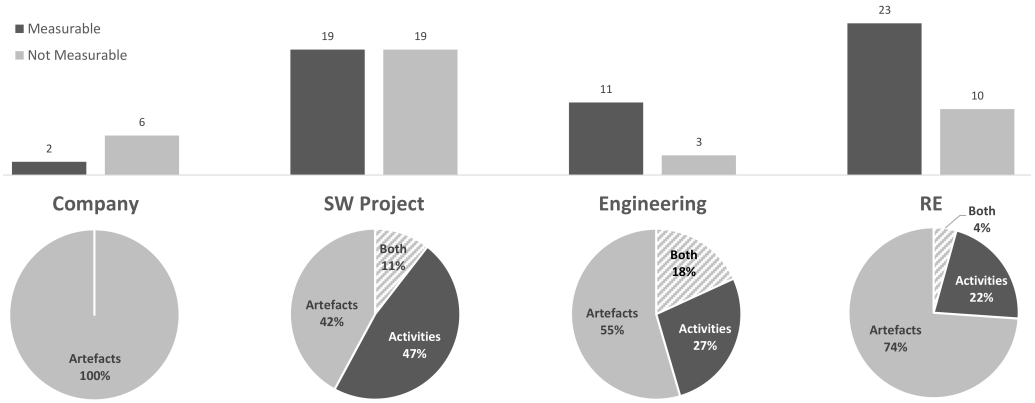

Figure 2: Measurability of variables regarding the dimensions *RE*, *Engineering*, *Project* and *Company*.

jects it can be measured. Fig. 2 illustrates the results per dimension.

Out of all 93 variables, we considered less than half (38 variables, 41%) not measurable at all. Considering the remaining 55 measurable variables, the majority was measurable exclusively on artefacts (33 variables, 60%), with measurability on activities coming second (17 variables, 31%). Only 5 variables (9%) were found to be measurable on both artefacts and activities.

Regarding RE phenomena only, out of 33 variables less than one-third were considered unmeasurable (10 variables, 30%). Unmeasurable variables comprise social phenomena, e.g. *Weak access to customer needs* and *Insufficient support by project lead*, as well as phenomena attributable to uncertainty and limited knowledge inherently present during RE, e.g., *Technically infeasible requirements* or *Implicit requirements not made explicit*. Measurability of RE phenomena was mostly considered possible exclusively on artefacts (17 variables, 74%) instead of activities (5 variables, 22%), with the requirements specification as the predominant artefact to measure RE phenomena, e.g., *Underspecified requirements*, *Instable requirements* or *Inconsistent object models*. Solely one variable (4%), namely *Informal (unpaid) changes during RE*, was considered to be measurable on both artefacts and activites.

DOI: http://dx.doi.org/10.1145/2601248.2601258

Regarding the variables in the engineering dimension, only 3 variables (21%) were considered unmeasurable, namely *Too complex solutions*, *Wrong design decisions* and the *Use of throw-away prototypes*. The remaining 11 variables (79%) were considered to be measurable on artefacts (6 variables, 55%, e.g. *Bugs and Defects*), activities (3 variables, 27%, e.g. *Increased discussion during implementation*) or both (6 variables, 18%, e.g. *Gold plating*).

Out of all variables in the project dimension, exactly half (19 variables, 50%) were considered measurable, with the following distribution: eight of them (42%) were classified as measurable on artefacts , nine on activities (42%), the remaining two variables (11%) were classified as measurable on both artefacts and activities.

Only two variables (25%) attributed to the company dimension, *No further improvement after acceptance* and *Communication flaws in teams*, were considered measurable, both on artefacts exclusively. However, the majority of variables (75%), e.g. *Customer dissatisfaction* and *Volatile domain*, were considered unmeasurable.

4.3 Actionability of Variables (RQ 3)

Actionability (cf. Sec. 3) denotes the ability to take immediate actions to significantly change a phenomenon. Fig. 3 illustrates the results of the classification.

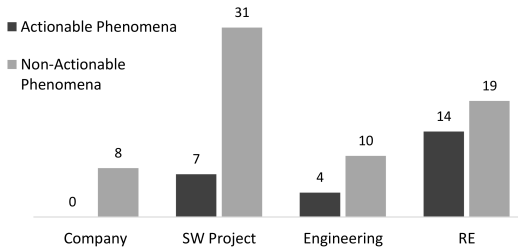

Figure 3: Actionability according to dimensions.

Overall, only a minority of variables (25, 27%) were considered actionable. Actionable variables are, e.g., *No validation* and *Underspecified requirements*. Those phenomena can even be avoided by introducing a specific RE process or a reference model for the artefacts that defines a notion of quality for the artefacts. *Increased process costs* and *Unavailability of customer*, in turn, are instances for variables considered inactionable.

The percentage of actionable variables (42%, compared to all variables within a dimension) is highest in RE. In contrast, 4 (29%) engineering variables and 7 (19%) project variables are actionable. None of the eight variables in the company dimension is actionable.

4.4 Reliability of Classifications

In order to assess the degree of the inter-rater agreement and, thus, the reliability of the classifications, we apply Cohen's κ-measure [3]. It is defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

in which P_o is the observed percentage agreement, and P_e is the expected probability of agreement among raters due to chance, based on marginal probabilities, i.e. the distribution of the actual selections of the raters, and complete statistical independence of raters. P_e estimates the proportion of times raters would agree if they guessed completely on every case and with probabilities that match the marginal proportions of the observed classifications.

The values of κ are constrained to the interval $[-1; +1]$. A κ value of one means perfect agreement, a κ value of zero means that agreement is equal to chance, and a κ value of negative one means perfect disagreement. For $0 \leq \kappa \leq 1$, literature suggests $\kappa > 0.60$ corresponds to a "good" or "substantial agreement", the range $0.21 - 0.60$ is interpreted as "fair" or "moderate" agreement in the majority of cases, while for the range $0 - 0.20$ we have "poor" agreement (see, e.g., [8]).

In case the distribution of ratings is skewed, the κ coefficient must be adjusted for prevalence, resulting in $2P_o - 1$. Tab. 3 reports the three mentioned agreement metrics. For the classification of the dimension, which was done by 2 raters, the Cohen's Kappa is computed. For the classifications of the measurability and the actionabilit, provided by 3 raters, the Fleiss [8] adjusted Kappa is computed.

We observe moderate agreement for the classification of dimensions and actionability. The agreement values on measurability are conservative, because we ignored the partial agreements (i.e. when a rater classified a variable as measurable on basis of both activities and artefacts and the other(s) only on basis of one of them). We report poor agreement in measurability for except for *SW project* and *RE*. The dimension *Engineering* has moderate agreement in measurability. We observe disagreeement for dimension *Company* both in measurability and actionability. Although only 8 nodes are affected, this reveals a diifculty for the raters to classifiy and interpret those variables.

Classification		P_o	$2P_o - 1$	Cohen's κ
Dimensions		0.60	0.20	0.43
Measurability	Overall	0.43	-0.15	0.20
	Company	0.38	-0.25	-0.13
	Engineering	0.52	0.05	0.34
	RE	0.42	-0.15	0.15
	SW project	0.39	-0.23	0.13
Actionability	Overall	0.67	0.34	0.34
	Company	0.50	0.00	-0.33
	Engineering	0.76	0.52	0.52
	RE	0.66	0.31	0.30
	SW project	0.68	0.37	0.34

Table 3: Reliability of classifications

5. CRITICAL REFLECTION

In the following, we critically reflect on our results by considering the possibilities and limitations for evidence-based RE research we draw from the extent to which variables are measurable and actionable in context of RE.

To this end, we first discuss the positive effects we see in the results, i.e. the measurability of RE phenomena within RE itself. In a second step, we critically discuss the negative conclusions we draw from our results. Those negative effects can be discussed, in turn, on two levels: we have variables that are barely measurable within RE or barely measurable at all, and we have variables where the measurements strongly depend on subjective interpretation, i.e. where the variables are strongly dependent on the particularities of a narrow ("local") socio-economic context with limited possibilities of generalisation. The latter is important to the more general, and to some extent philosophical,

DOI: http://dx.doi.org/10.1145/2601248.2601258

question to which extent the variables can eventually serve at all to build and evaluate contributions to RE that are not dependent on subjectivity (i.e. oracles).

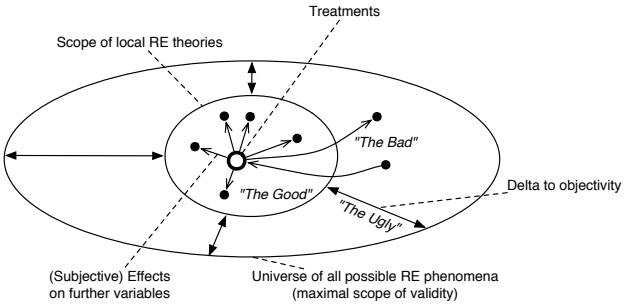

Figure 4: The Good, the Bad, and the Ugly: Possibilities and limitations in evidence-based RE research.

Figure 4 summarises the perspectives we take in the following while categorising them according to the possibilities we conclude for evidence-based RE research.

5.1 The Good: Measurability in RE Context

We see those dependent variables that allow us to reliably measure the effects of treatments in a specific experimental context with a low degree of subjectivity[2]as most valuable ("The Good", Fig. 4) as they allow us, for example, to accurately test the sensitivity of applying an RE method in a socio-economic context.

We discovered that 59 % of all revealed variables can be measured while 69 % of those variables can be measured on basis of artefacts. For those variables that origin in context of RE, we discovered that 78 % of the variables are measurable on basis of the artefacts with a high percentage of actionability compared to the variables allocated to other dimensions. One implication we draw is that when applying a treatment to RE, we can rely on independent and comparable measurements on basis of artefacts rather than relying on measurements on basis of software engineering activities. This understanding results from our experiences that by taking artefact reference models, we can define a certain external notion of artefact quality to which we can compare the actual results of an experiment in a standardised manner [15].

Apart from the possibilities we see in the means for measuring the variables is that a substantial amount of variables is allocated to RE. This enables to shorten the empirical cycle when conducting, for example, case study research as we do not necessarily have to take into account subjects and objects associated with the whole development process. For example, we can investigate the effects of a RE specification method by directly investigating the quality of the created RE artefacts while isolating investigations of the effects on other development activities, e.g. on reviews within analytical quality assurance tasks.

From the perspective of evidence-based research, this means that we are able to set up an experiment or a case study while focussing on RE only, while neglecting a large extent of phenomena relevant to further dimensions, but still preserving potential effects on them. The latter is reflected by a set of discovered variables having impacts on further variables that would demand for longitudinal studies (e.g., incomplete requirements causing change requests).

For researchers, this supports the accuracy in the planning of experimental settings. Practitioners can already use those results, for example, to calibrate their quality assurance techniques.

5.2 The Bad: Limitations in RE Context

One direct implication of the foregoing discussion is that, in contrast to the given measurability in RE, we could also reveal a set of cause-effect relationships between variables that are barely measurable or not measurable in a local, context-specific RE setting. We consider those variables to limit the possibilities in evidence-based RE research ("The Bad", Fig. 4). That is:

1. 73 % of the variables are not actionable, let alone as they result from circumstances that are not visible when setting up a project (e.g. the effects the relationship to customers has on the quality of the RE artefacts), or they are
2. not measurable (41 %) or not objectively measurable limiting the internal and the external validity of measurements (e.g. the customer satisfaction), or
3. they are measurable, but they have relationships with variables not within RE making measurements and especially their interpretation difficult (e.g. incomplete requirements with a strong causal relation to 4 variables in RE, but also to 8 variables in the project dimension and even one in the overall company dimension), or, finally, they
4. have no direct relation to RE (except over transitive and, thus, not empirically resilient relationships).

Therefore, while we see good chances to use a substantial amount of variables for measurements in RE, we still have a large extent of variables that imply the need of further investigation on (1) means for accurate measurements and (2) extension of the variables themselves. The latter considers the need to better understand the further dimensions via longitudinal studies and replication studies as well to strengthen the external validity of those variables that by now remain underrepresented (see also Sect. 7.1 on the threats to validity).

5.3 The Ugly: No RE Oracle in Sight!

Apart from problems in the measurability of certain variables discussed in the previous section, we consider another problem to affect the generalisability of empirical results relying on the obtained variables. That is, the notion of objectivity[3] and, thus, external validity is weak negatively affecting the extent to which we are eventually able to obtain RE oracles in general, i.e. the possibility to eventually establish universally valid RE theories is, in our opinion, low ("The Ugly", Fig. 4).

The reason is that especially in RE, a large extent of variables usable in experimental settings strongly relates to a measurements where experiences, the expertise, and the ex-

[2]We speak of a high degree of "subjectivity" when the outcome of a measurement is dependent on the particularities and attitude of the involved subjects.

[3]In its essence, objectivity considers that a treatment results always in the same effects when applying it independently to a population.

DOI: http://dx.doi.org/10.1145/2601248.2601258

pectations of the subjects involved strongly affect the sensitivity of RE contributions tested in socio-economic contexts. For example, when testing the efficiency of an RE elicitation method, the results will always depend on beliefs and the judgment of those who apply the method. In consequence, one might assume that there is no such thing as universal truth and that an oracle for RE will never be at our disposal, because the effects of a treatment will always depend on the subjective view of human beings involved in the context. An alternative view is that it would be inherently hard to identify and eliminate all confounding factors.

Even if ignoring that we still have many unknown variables in our results or variables where the measurements rely on subjective interpretation, those variables we could already obtain negatively affect the possibility of generalisation, because:

1. most variables important to RE have complex and often transitive cause-effect relationships, and
2. they have a low degree of measurability.

We can see, for example, that underspecified or unmeasurable requirements impact the effort spent for reviews. We now could investigate more precisely those effects by conducting a longitudinal study and come, for example, to the conclusion that one particular specification method applied to RE has the effect of leading to more precise RE artefacts and a decreasing review effort. However, the problem in such an investigation is of more general nature as the decreased effort could be also caused by variables not included in the experimental setting and side-effects not taken into account. One reason why a potential correlation will, in our opinion, never imply causation is that we doubt that it is possible to build yet such a system of dependent variables that is able to capture all possible facets of a project ecosystem.

The results in their current state already reveal the complexity in the dependent variables; for example, change requests are already caused by 9 different variables, moving targets already affects 10 variables.

However, assuming that an RE theory is always something relative, we can at least use the variables to establish and calibrate experimental settings as long as we rely on those aspects we can measure in RE in a standardised manner (e.g. on basis of artefacts, see Sect. 5.1) or by explicitly opting for subjectivity, e.g. by gathering expert opinions. The latter makes clear the necessity to explicitly and accurately characterise experimental contexts (subjects, background, expertise etc.).

Finally, although one might doubt the possibility to generally obtain full external validity, we can (and should) increase the external validity of experimental results until reaching a certain saturation via a tool already at our disposal: replication studies [12].

6. RESEARCH IMPLICATIONS

Starting from our results, we picture three areas of implications for which we encourage researchers and practitioners to foster the necessary discussions, namely (i) general principles in evidence-based RE research, (ii) RE methodologies and (iii) RE quality management.

6.1 Evidence-based RE Research in General

Implication we can draw from our results on evidence-based RE research mainly result from the negative effects we discussed in the previous section. Simplified, we have

1. a large extent of variables that are hard to measure
2. a limited understanding on the context surrounding the particularities in RE.

One might now argue that only the limited set of measurable variables with a low degree of dependencies are suitable to conduct empirical studies. However, we strongly believe that even if avoiding a pragmatic view on the results presented in previous sections, we can already make use of all variables allocated to the context of the RE dimension while explicitly opting for those for which the measurements is inherently subjective. That is, the results increase the awareness of those variables that demand for expert judgement, thus, allowing us to calibrate the experimental setting, e.g., via survey research. To increase the reliability of the results and to tackle the problem of a limited understanding on the context surrounding the particularities in RE, we belief that we should especially value more independent (confirmatory) replication studies [12].

Apart from the general implications on the various types of empirical studies, we draw the need to conduct more curiosity-driven studies in RE to reveal more variables and strengthen the confidence in those variables we already could define. This should support a better understanding on the general phenomena in RE necessary to, for example, infer proper improvement goals or general characteristics suitable to tailor RE methodologies to practical environments. As a matter of fact, this extension of the results is already in scope via the globally distributed replications as part of the NaPiRE endeavour.

Nevertheless, the variables presented already serve practitioners and researchers to calibrate their improvement goals and the metrics used in evaluation research and is in scope of the following section.

6.2 Research on RE Methodologies

When investigating the area of RE methodologies, i.e. benefits and needs when relying on certain RE methods or whole software process models, we too often rely on opportunistically chosen metrics for measurements that might be important to that particularly envisioned socio-economic context while for others, it might be not. To test the sensitivity of a method in a context, we therefore believe in the benefits of relying on measurements for those variables stated with a high number of occurrences in our result set. The reasons is that those variables seem important to a broader range of practitioners, thus, they already imply an increase in external validity for the corresponding improvement goals. Furthermore, by relying on commonly accepted practical problems to infer improvement goals, we allow for accurate objectivism via independent replication studies.

Exemplary improvement goals we can already infer from our results are:

1. Increase flexibility in the RE process (reflecting exemplary problems like moving targets, time boxing, or gold plating)
2. Increase syntactic quality in RE artefacts (reflecting exemplary problems like inconsistent requirements, terminological problems)
3. Support precise terminology and communication
4. Support consistency and traceability
5. Support testability of requirements
6. Make explicit the particularities of the application domain

DOI: http://dx.doi.org/10.1145/2601248.2601258

7. Increase of semantic quality of the RE artefacts (reflecting exemplary problems like incomplete / hidden requirements)

Those improvement goals already show the diversity of treatments we can test while still being in scope of practically relevant problems. Some goals envision whole methodologies and the way of working (activities), others envision the quality of the artefacts and the specified requirements. The nature of the measurements for the corresponding variables furthermore reveal first implications on the necessary type of studies. While those variables relying on subjective measurements should be envisioned via, for example, technical action research case studies [22] and study types to gather subjective expert judgement (e.g., survey research, or grounded theory) [1], other more objectively measurable variables could be in scope of controlled environments. The cause-effect relationships of the variables can be finally used to design longitudinal studies, e.g. to test the effects of a treatment on the general project quality (see also the discussion in Sect. 5).

Another field of application considers the validation of previously conducted studies and the calibration of the metrics used for further studies. For example, we have already tested the benefits and shortcomings of applying artefact orientation to RE in various experiments and case studies.[14, 13]. We could also confirm the value of replication studies as those studies confirmed trends, e.g. the support of consistency and a clear terminology. While most of the variables we used in our studies match indeed the stated improvement goals, there are some variables we plan to remove for further replication as they show to have too many dependencies and uncontrollable side-effects (e.g., the efficiency of the tested approaches).

We strongly believe that we have the same effects when testing the practical impact of single methods in evaluation research. That is, if practitioners and researchers test the sensitivity of applying single methods (e.g., for requirements elicitation), they can rely on our presented variables following corresponding study types while isolating those variables that are inherently threatened in their validity.

6.3 Research on RE Quality Management

In the following, we discuss research implications on quality management in RE, comprising both quality assurance (assessment) and quality control (correcting actions).

Several indicators suggest the further investigation of metrics in RE quality management. While the results of Sec. 4.2 suggest both a substantial basis (23 variables) and the most promising ratio (70%) of measurable variables, the inter-rater agreement for RE measurability ($\kappa = 0.15$) and actionability ($\kappa = 0.30$) was rather low, especially in contrast to the Engineering dimension ($\kappa = 34$ resp. 52). This suggests that for RE, we are still lacking knowledge of how to assess and control practical problems using metrics. However, there are immediate benefits: on the one hand, several phenomena (e.g. *Inconsistent* and *Incomplete/Hidden requirements*) considered measurable but not actionable may be refined into precise defects and measures, potentially yielding actionability. On the other hand, for those phenomena considered actionable but not measurable (R05,R08,RP11; e.g., *Underspecified requirements*), the discovery of adequate metrics could improve quality management in RE notably. In particular, artefact-based metrics seem most promising, because (i) 78% of RE phenomena can be measured on basis of artefacts, and (ii) artefact-based reference models can be leveraged to standardised measurements to obtain comparability.

Indeed, the relation of many identified RE problems, e.g., *Uncertainty in RE* and *Weak access to customer needs*, to the more traditional notions of RE quality in general and their manifestations and impacts of requirement specifications in terms of quality, remains unclear. Bridging this gap between the observations and expectations of practitioners and scientific RE quality models would benefit both: Practitioners would have easier access to scientific RE quality management methods and tools, while the scientific community could profit from an empirically-grounded notion of quality, evaluating and revising existing models.

7. CONCLUSION AND FUTURE WORK

In this paper, we contributed a set of dependent RE variables classified according to the dimensions defined by Gorschek et al. [9]. We showed to what extent the variables are measurable and actionable. Based on this classification, we critically reflected on the results from the perspective of evidence-based research in RE and draw, as a second step, direct implications for RE research.

Our results showed that out of 93 variables, 33 have their origin in RE. Furthermore, we discovered that 59 % of all revealed variables can be measured while 69 % of those variables can be measured on basis of artefacts. In RE, even 78 % of the variables are measurable on basis of the artefacts. This strengthens our confidence in the possibility to accurately set up an experiment or a case study while focussing on RE only, and while neglecting a large extent of phenomena relevant to further dimensions, but still preserving potential effects on them. However, we could also show a substantial amount of variables outside RE having complex dependencies. We critically discussed the implications on evidence-based RE research and, for example, on the possibility obtain objectivity in experimental settings.

Further implications we draw were on the RE research itself, i.e. we showed

1. the need to conduct longitudinal studies and confirmatory replication studies for RE,
2. a first set of improvement goals suitable to calibrate experimental settings in research on RE methodologies, and
3. the implications on research on RE quality management.

7.1 Limitations and Threats to Validity

A threat to the validity is given by the classification itself given that the area of metrics and measurements is inherently complex and multi-facetted. We minimised this threat via research triangulation. However, Tab. 3 shows mainly "moderate" agreements and even disagreement for variables in the *Company* dimension. We contrasted the moderate reliability by solving conflicts with specific reconciliation meetings, in which the three raters concurred by discussion on an unique classification.

The biggest threat to the validity of our results remains, however, the incompleteness of the variables. We have, for example, revealed small coherent sets of variables independent of RE and often stated in a limited number of occurrences. However, those isolated groups of dependent vari-

DOI: http://dx.doi.org/10.1145/2601248.2601258

ables make also clear the necessity (and possibilities) of further investigations which we actively discussed. One explanation we have is that the focus when eliciting the variables within NaPiRE was on RE-specific variables and their causes and effects leaving open an understanding about the phenomena within the other dimensions. The replication and, thus, the extension of the results also to other dimensions is in scope of the NaPiRE endeavour whereas our results already allow to foster the discussion on the implications on research in RE.

7.2 Future Work

We are currently replicating the globally distributed surveys on RE in various countries We plan to use the results to extend and validate the classification we provided with this paper.

8. REFERENCES

[1] S. Adolph, W. Hall, and P. Kruchten. Using Grounded Theory to study the Experience of Software Development. *Journal of Empirical Software Engineering*, 16(4):487–513, 2011.

[2] Cheng, B.H.C. and Atlee, J.M. Research Directions in Requirements Engineering. In *Future of Software Engineering (FOSE'07)*, pages 285–303. IEEE Computer Society, 2007.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[4] N. Condori-Fernández, M. Daneva, and R. Wieringa. Preliminary Survey on Empirical Research Practices in Requirements Engineering. Technical Report TR-CTIT-12-10, University of Twente, Centre for Telematics and Information Technology (CTIT), 2012.

[5] D. Damian and J. Chisan. An Empirical Study of the Complex Relationships between Requirements Engineering Processes and other Processes that lead to Payoffs in Productivity, Quality, and Risk Management. *IEEE Transactions on Software Engineering*, 32(7):433–453, 2006.

[6] K. El Enam and N. Madhavji. A Field Study of Requirements Engineering Practices in Information Systems Development. In *Proceedings o the 2nd IEEE International Symposium on Requirements Engineering*, pages 68–80. IEEE Computer Society, 1995.

[7] J. Eveleens and T. Verhoef. The Rise and Fall of the Chaos Report Figures. *IEEE Software*, 27(1):30–36, 2010.

[8] J. L. Fleiss. *Statistical Methods for Rates and Proportions*. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, second edition, 1981.

[9] Gorschek, T. and Davis, A.M. Requirements Engineering: In Search of the dependent Variables. *Information and Software Technology*, 50:67–75, 2007.

[10] T. Hall, S. Beecham, and A. Rainer. Requirements problems in twelve software companies: an empirical analysis. *Empirical Software Engineering*, 8(1):7–42, 2003.

[11] P. Hsia, A. Davis, and D. Kung. Status report: Requirements engineering. *IEEE Software*, 10(6):75–79, 1993.

[12] N. Juristo and S. Vegas. Using differences among replications of software engineering experiments to gain knowledge, 2009. Invited Talk from the International Conference on Empirical Software Engineering and Measurement (ESEM).

[13] M. Kuhrmann, D. Méndez Fernández, and A. Knapp. Who Cares About Software Process Modelling? A First Investigation About the Perceived Value of Process Engineering and Process Consumption. In *PROFES'14*, pages 138–152. Springer, 2013.

[14] D. Méndez Fernández, K. Lochmann, B. Penzenstadler, and S. Wagner. A Case Study on the Application of an Artefact-Based Requirements Engineering Approach. In *EASE'11*, pages 104–113. Institution of Engineering and Technology (IET), 2011.

[15] D. Méndez Fernández, B. Penzenstadler, M. Kuhrmann, and M. Broy. A Meta Model for Artefact-Orientation: Fundamentals and Lessons Learned in Requirements Engineering. In D. Petriu, N. Rouquette, and O. Haugen, editors, *MoDELS'10*, volume 6395, pages 183–197. Springer-Verlag Berlin Heidelberg, 2010.

[16] D. Méndez Fernández and S. Wagner. Naming the Pain in Requirements Engineering: Design of a global Family of Surveys and first Results from Germany. In *EASE'13*, pages 183–194. ACM Press, 2013.

[17] D. Méndez Fernández, S. Wagner, K. Lochmann, A. Baumann, and H. de Carne. Field Study on Requirements Engineering: Investigation of Artefacts, Project Parameters, and Execution Strategies. *Information and Software Technology*, 54(2):162–178, 2012.

[18] Méndez Fernández, D. and Wagner, S. Naming the Pain in Requirements Engineering. *Information and Software Technology*, Currently in Revision, Author Version available under: `http://www4.in.tum.de/~mendezfe/openspace/NaPiRE/INFSOF-S-13-00428.pdf`, 2013.

[19] A. Meneely, B. Smith, and L. Williams. Validating software metrics: A spectrum of philosophies. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 21(4):24, 2012.

[20] U. Nikula, J. Sajaniemi, and H. Kälviäinen. A State-of-the-practice Survey on Requirements Engineering in Small-and Medium-sized Enterprises. Research Report 951-764-431-0, Telecom Business Research Center Lappeenranta, 2000.

[21] B. Nuseibeh and S. Easterbrook. Requirements Engineering: A Roadmap. In *Proceedings of the Conference on the Future of Software Engineering*, pages 35–46, New York, NY, USA, 2000. ACM.

[22] R. Wieringa and M. Aycse. Technical action research as a validation method in information systems design science. In *Proceedings of the 7th international conference on Design Science Research in Information Systems: advances in theory and practice*, pages 220–238, 2012.